Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 92 (2021) 104831

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Molecular epidemiology analysis of early variants of SARS-CoV-2 reveals the potential impact of mutations P504L and Y541C (NSP13) in the clinical COVID-19 outcomes

Canhui Cao [a,b], Liang He [a,c], Yuan Tian [a,c], Yu Qin [a,c], Haiyin Sun [a,c], Wencheng Ding [a,c], Lingli Gui [d,**], Peng Wu [a,c,*]

[a] Cancer Biology Research Center (Key Laboratory of the Ministry of Education), Tongji Medical College, Tongji Hospital, Huazhong University of Science and Technology, Wuhan 430030, China
[b] Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Shenzhen Hospital, Shenzhen, Guangdong 518036, China
[c] Department of Gynecologic Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[d] Department of Anesthesiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2 strains
Genetic variations
Amino acid variations
ORF1ab

ABSTRACT

Since severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused global pandemic with alarming speed, comprehensively analyzing the mutation and evolution of early SARS-CoV-2 strains contributes to detect and prevent such virus. Here, we explored 1962 high-quality genomes of early SARS-CoV-2 strains obtained from 42 countries before April 2020. The changing trends of genetic variations in SARS-CoV-2 strains over time and country were subsequently identified. In addition, viral genotype mapping and phylogenetic analysis were performed to identify the variation features of SARS-CoV-2. Results showed that 57.89% of genetic variations involved in ORF1ab, most of which (68.85%) were nonsynonymous. Haplotype maps and phylogenetic tree analysis showed that amino acid variations in ORF1ab (p.5828P > L and p.5865Y > C, also NSP13: P504L and NSP13: Y541C) were the important characteristics of such clade. Furthermore, these variants showed more significant aggregation in the United States ($P = 2.92E-66$, 95%) than in Australia or Canada, especially in strains from Washington State ($P = 1.56E-23$, 77.65%). Further analysis demonstrated that the report date of the variants was associated with the date of increased infections and the date of recovery and fatality rate change in the United States. More importantly, the fatality rate in Washington State was higher (4.13%) and showed poorer outcomes ($P = 4.12E-21$ in fatality rate, $P = 3.64E-29$ in death and recovered cases) than found in other states containing a small proportion of strains with such variants. Using sequence alignment, we found that variations at the 504 and 541 sites had functional effects on NSP13. In this study, we comprehensively analyzed genetic variations in SARS-CoV-2, gaining insights into amino acid variations in ORF1ab and COVID-19 outcomes.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes coronavirus disease 2019 (COVID-19), is a novel RNA virus from the Coronaviridae family (Wu et al., 2020). It is related to SARS-like coronaviruses found in bats (e.g., bat-SL-CoVZC45 and bat-SL-CoVZXC21) (Hu et al., 2018; Lu et al., 2020), and shares 79% genetic similarity to SARS-CoV and ~ 50% genetic similarity to Middle East respiratory syndrome (MERS)-CoV (Lu et al., 2020). Notably, SARS-CoV-2 binds to the angiotensin-converting enzyme 2 (ACE2) receptor with a similar receptor-binding domain as SARS-CoV (Lu et al., 2020; Zhou et al., 2020b). In late December 2019, the first cluster of COVID-19 patients was confirmed and reported (Zhu et al., 2020). By 28 January 2020, infections had reached 14,900 (Lu et al., 2020; Zhou et al., 2020b). Based on rapid increase in the number of infections, COVID-19 had become a global pandemic with alarming speed. As of 3 March 2021, there have been more than 113 million infections and more than 2.5 million deaths worldwide. Notably, the United States has recorded

* Corresponding author at: Cancer Biology Research Center (Key Laboratory of the Ministry of Education), Tongji Medical College, Tongji Hospital, Huazhong University of Science and Technology, Wuhan 430030, China.
** Corresponding author.
E-mail addresses: gui_lingli@hotmail.com (L. Gui), pengwu8626@126.com (P. Wu).

https://doi.org/10.1016/j.meegid.2021.104831
Received 24 December 2020; Received in revised form 8 March 2021; Accepted 28 March 2021
Available online 31 March 2021

28.5 million confirmed infections and 0.5 million deaths due to COVID-19, accounting for 25.08% of global infections and 20.36% of global deaths, thus making it one of the most severely affected countries. However, the reasons for the rapid spread of this novel virus and the pathogenesis of infections remain to be determined.

Recent studies focused on the mutation of Spike protein and showed that variants carrying D614G have become the most prevalent worldwide, suggesting the fitness advantage for SARS-CoV-2 (Canhui et al., 2020; Korber et al., 2020). However, variations in other genes beyond the Spike protein might be important for the evolution of this virus. SARS-CoV-2 was indicated to evolve into at least three phylogenetic groups, characterized by positive selection from ORF1ab, ORF3a, and ORF8. Of note, for the first time, Velazquez-Salinas et al. identified the potential relevance of amino acid Y5865C in ORF1ab, showing that this residue is experimenting directional selection. Also, the increased evolutionary rate of ORF10 was identified by Velazquez-Salinas et al. (Velazquez-Salinas et al., 2020). Moreover, the exoribonuclease (ExoN) of NSP14 knockout mutant assays showed the replication roles of enzymatic activity in MERS-CoV and SARS-CoV-2 (Ogando et al., 2020). In addition to analyzing the genetic diversity of SARS-CoV-2, the understanding of pathogen lineage could help scientists effectively target interventions, track variants, and improve interpretations of pathogenesis data (du Plessis et al., 2021). It was reported that more than 1000 lineages were being established before the national lockdown in the United Kingdom (du Plessis et al., 2021). At the phylogeny of SARS-CoV-2, strains were identified as lineage A, B, and C according to variants, and the variant carrying the Y5865C mutation was associated with the lineage S or 19 B (Velazquez-Salinas et al., 2020).

In the correlation analysis, variants of S 614G was associated with case fatality rates (CRF) and median CFR in 12 countries (Becerra-Flores and Cardozo, 2020), and variants ORF1ab 4715 L and S 614G were reported correlated with fatality rates in 28 countries and 17 states of the United States (Toyoshima et al., 2020). It is still unclear whether the different mortality rates or transmission rates observed in different regions may be the consequences of the differences in clade virulence (Mercatelli and Giorgi, 2020). ORF1 accounts for about two-thirds of the whole genome and encodes two polyproteins, i.e., pp1a (approximately 486 kD) and pp1ab (approximately 790 kD) (Hilgenfeld and Peiris, 2013), which are processed by two viral cysteine proteases, i.e., papain-like protease (PLpro, Nsp3 domain) and main protease (Mpro or 3CLpro, Nsp5), into 15 or 16 NSPs. Most of these Nsps are involved in the transcription or replication of viral genomes (Sawicki et al., 2007).

In this study, we explored 1962 genomes of SARS-CoV-2 strains obtained from 42 countries to analyze the correlations between genetic variation and disease outcomes of COVID-19. We then identified the changing trends of genetic variations in SARS-CoV-2 strains over time and by country. We also performed viral genotype mapping and phylogenetic tree analysis to determine the variation features of SARS-CoV-2. Based on the infection, fatality, and recovery rates, as well as dynamic curves for the emergence of genome variations, in different countries, we identified amino acid variations in ORF1ab at the 5828 and 5865 loci (NSP13: P504L and NSP13: Y541C) and gained insight into COVID-19 outcomes in the United States that contained different proportions of strains with these ORF1ab variations.

2. Materials and methods

2.1. Genetic variations in SARS-CoV-2 strains

Genetic variations in SARS-CoV-2 strains were determined based on data obtained from 2019nCoVR (v2.1) (https://bigd.big.ac.cn/ncov) (Zhao et al., 2020; Song et al., 2020) of the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB)/ Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences. We explored 1962 viral strains with complete genomes sampled from 13 December 2019 to 21 March 2020. To demonstrate genetic variations in SARS-CoV-2 strains, we performed the variation frequency analysis of whole genomes over time and by country and then constructed dynamic curves for genetic variations across different countries, as well as viral genotyping maps of haplotypes by country based on genetic variation of high-quality genomic sequences and phylogenetic trees across different countries based on genome variation using MEGA. All results were performed with default settings in 2019nCoVR (v2.1).

2.2. Genetic variation heatmap

Genetic variation heatmaps of SARS-CoV-2 strains were generated by the online tool (Spatiotemporal dynamics) in 2019nCoVR (v2.1) (https://bigd.big.ac.cn/ncov) (Zhao et al., 2020, Song et al., 2020). By calculating the frequency of mutation sites of SARS-CoV-2 strains at each time point, the dynamic trend over time is displayed in the form of a genetic variation heatmap. By calculating the frequency of mutation sites in each country, the dynamic trend in different countries was displayed in the form of a genetic variation heatmap.

2.3. Nonsynonymous and synonymous variation in SARS-CoV-2 strains

Amino acid variations in SARS-CoV-2 strains corresponded to genetic variation data. The mutation number of virus regions (5'UTR, ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF8, N, ORF10) represented the number of strains with genetic variation at each site. The variation annotation reference was NCBI: txid2697049, which was submitted by the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China (Wu et al., 2020). The full SARS-CoV-2 proteome was based on the NCBI reference sequence (NC_045512), with GenBank entry MN908947. The mutation density of each virus region was calculated by dividing the mutation number of strains by the length of each region (bp).

2.4. Genome variations in ORF1ab at 17747 and 17858

Genetic variations in ORF1ab: 17858 (A- > G) and ORF1ab: 17747 (C- > T) corresponded to amino acid variations in ORF1ab: p.5865Y > C and ORF1ab: p.5828P > L, or NSP13: P504L and NSP13: Y541C. In the protein sequence, the amino acid of ORF1ab at site 5828 P (proline) mutated into L (leucine) and at site 5865 Y (tyrosine) mutated into C (cystine).

2.5. Viral genotyping maps and phylogenetic tree of SARS-CoV-2 strains

Viral genotyping maps were used to perform haplotype analysis across countries based on 2019nCoVR (v2.1) (https://bigd.big.ac.cn/ncov) (Zhao et al., 2020; Song et al., 2020). We used 1962 viral strains for genetic variation analysis and 1211 strains for haplotype maps (Fig. S1). Strain numbers used for haplotype and genetic variation analyses were different due to the asynchronous operation process of these two results in the dataset. We used the latest 2250 strains of the virus from 13 December 2019 to 26 March 2020, identifying 1210 haplotypes in total. Haplotype network maps were used to demonstrate the genetic distance and evolutionary relationships among different haplotypes. Root and leaf nodes indicated the direction of SARS-CoV-2 variants. The phylogenetic tree was built using 2019nCoVR with default settings. The phylogenetic tree was analyzed using the Molecular Evolutionary Genetics Analysis (MEGA) tool, with a scale of 0.0001. Two phylogenetic treemaps based on ORF1ab: p.5828P > L and ORF1ab: p.5865Y > C were download from Nextstrain (https://nextstrain.org/ncov) (Hadfield et al., 2018).

2.6. Sequence alignments of SARS-CoV-2 and other coronaviruses

The amino acid sequences of ORF1ab from SARS-CoV-2 and other coronaviruses (i.e., *Pipistrellus* bat coronavirus HKU5, Bat coronavirus

BM48–31/BGR/2008, Porcine epidemic diarrhea virus, Canine respiratory coronavirus, Ferret coronavirus, Beta coronavirus *Erinaceus*/VMC/DEU, Human coronavirus OC43, Bat coronavirus_QBP43288, Hedgehog coronavirus 1, *Tylonycteris* bat coronavirus HKU5, *Tylonycteris* bat coronavirus HKU4, BatCoV RaTG13, MERS, SARS, bat-SL-CoVZC45, and bat-SL-CoVZXC21) were downloaded from the Protein Database of NCBI (https://www.ncbi.nlm.nih.gov/protein).

2.7. COVID-19 case collections

The COVID-19 infection, mortality, and recovery rates in the United States were collected from the Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/) and virusncov (https://virusncov.com/covid-statistics/usa), and included confirmed and presumptive positive cases reported to the CDC since 22 January 2020, not including cases repatriated to the United States from Wuhan (China) or Japan. COVID-19 cases, mortality rates, and recovery rates of other countries were collected from the World Health Organization (WHO) (https://www.who.int/).

2.8. Protein secondary structure predictions

The α-helix, β-sheet, and β-turn structures were used to represent protein secondary structures. Protein secondary structure prediction of ORF1ab (QHD43415.1) and the ORF1ab variant (p.5865Y > C and p.5828P > L, also NSP13: P504L and NSP13: Y541C) was performed based on the Chou-Fasman algorithm (Chou and Fasman, 1974). We used the amino acid sequence of ORF1ab (QHD43415.1) and the ORF1ab variant (QHD43415.1: p.5865Y > C and p.5828P > L) to compare differences between the two ORF1ab proteins.

2.9. Functional effect prediction of ORF1ab variation

Polymorphism Phenotyping v2 (PolyPhen-2) is a tool for predicting the possible impact of the amino acid substitution or indel on the structure and function of a protein (Adzhubei et al., 2010). Protein Variation Effect Analyzer (PROVEAN) is also a tool for predicting the possible impact of an amino acid substitution or indel on the biological function of a protein (Choi et al., 2012; Choi and Chan, 2015). PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/index.shtml) and PROVEAN (v1.1) (http://provean.jcvi.org/index.php) were used to predict whether the variation in NSP13: P504L and NSP13: Y541C affected the protein function of NSP13. We used the amino acid sequence of the variants for the query. In PolyPhen-2, the HumDiv value was used to evaluate rare alleles and dense mapping of regions and to analyze natural selection, whereas the HumVar value was used to help diagnose Mendelian disease. The closer the value (HumDiv or HumVar) is to 1.0, the greater the effect on the protein function, and the closer the value (HumDiv or HumVar) is to 0, the less the effect on the protein function. Score curves (specificity and sensitivity) were drawn according to PROVEAN v1.1 (Choi et al., 2012). Score > −2.5 was classified as a neutral effect, and score < −2.5 was classified as deleterious effect.

2.10. Protein modeling

Homology modeling protein of ORF1ab and the variation of ORF1ab (P5828L and Y5865C, NSP13: P504L and NSP13: Y541C) of SARS-CoV-2 were performed by SWISS-MODEL Server (Bienert et al., 2017; Waterhouse et al., 2018) (https://swissmodel.expasy.org/interactive). The global quality estimate included QMEAN (Studer et al., 2021) (a composite estimator based on different geometrical properties), Cβ, all-atom, solvation, and torsion.

2.11. Statistical analysis

Data were displayed as the number of amino acid variations (synonymous or nonsynonymous), infection cases, mortality cases, and recovery cases. Statistical analyses were performed using SPSS software and interpreted by the Chi-squared test. A *P*-value of <0.05 was considered to show statistical significance.

3. Results

3.1. Landscape of genetic variations in SARS-CoV-2 strains

Since there was a moderate increasing trend after April, we tended to analyze the correlations between genetic variation and disease outcomes of COVID-19 in early strains before April (Fig. S1). We analyzed 1962 high-quality genomes of early viral strains obtained between 23 December 2019 and 21 March 2020 from the 2019nCoVR dataset. We first performed variation frequency integration for whole genomes of SARS-CoV-2 strains over time. We found 1660 sites with genetic variations, including several regions with a significant number of variations, i.e., ORF1ab (961) and S (162). Notably, variation frequency was markedly enriched from 27 February 2020 (Fig. 1A). Variation heatmap of countries demonstrated the composition of several significant genetic variation sites, including variation in ORF1ab (8782C- > T, NSP4:S76S, synonymous), S (23403 A- > G, S:D614G, missense), ORF1ab (17858 A- > G, NSP13:Y541C, missense), and ORF1ab (17747C- > T, NSP13:P504L, missense) (Fig. 1B).

We next conducted variation annotation of all genome variations according to the NCBI reference sequence NC_045512, with GenBank entry MN908947. In general, we found 1458 genome variations in coding regions of SARS-CoV-2, 69.48% of which were nonsynonymous. In addition, 66.05% (963/1458) of genetic variations were distributed in ORF1ab, of which 68.85% (663/963) were found to be nonsynonymous (Fig. 1C, Supplementary Table S1). As the length of ORF1ab occupied more than two-thirds of the whole genome, the mutation rate of each gene region should be determined in consideration of gene length. We calculated the mutation density of each gene via dividing the mutation number of each gene region by the length of each SARS-CoV-2 gene and found that the ORF10 region had a higher mutation rate (Fig. S3).

3.2. Characteristics of genetic variations in SARS-CoV-2 strains

To further map the genetic variations in coding regions of early SARS-CoV-2 strains, we displayed genetic variations within the whole genomes. The top variation sites were S: 23403 (S:D614G, 597 strains), ORF1ab: 3037 (NSP3:F106F, 595 strains), ORF1ab: 14408 (NSP12b:P314L, 594 strains), 5'UTR: 241 (592 strains), ORF1ab: 8782 (NSP4:S76S, 423 strains), ORF8: 28144 (ORF8:L84S, 423 strains), ORF1ab: 18060 (NSP14:L7L, 292 strains), ORF1ab: 17858 (NSP13:Y541C, 286 strains), and ORF1ab: 17747 (NSP13:P504L, 285 strains). Among them, 5'UTR was the untranslated region, and the variations in ORF1ab: 3037 (NSP3:F106F), ORF1ab: 8782 (NSP4:S76S), and ORF1ab: 18060 (NSP14:L7L) were synonymous (Fig. 2A). We also found that the average variation frequency of coding regions was 5.48 (ranging from 1 to 597) and of noncoding regions was 6.96 (ranging from 1 to 592).

To identify changes in genetic variations over time and by country, we constructed variation dynamic curves. Results showed that genetic variations in United States strains were not only responsible for the top nonsynonymous variations in the ORF1ab, S, and ORF8 regions but also were responsible for the top synonymous variations in the ORF1ab: 8782 and ORF1ab: 3037 sites. Furthermore, the occurrence date of the genetic variations in the United States strains contributed to the temporal trend of variations observed on 27 February 2020 (Fig. 2B).

3.3. Variation features of SARS-CoV-2 strains

Due to their extremely high mutation rates, short generation time, and large populations, viruses can rapidly develop viral quasispecies in

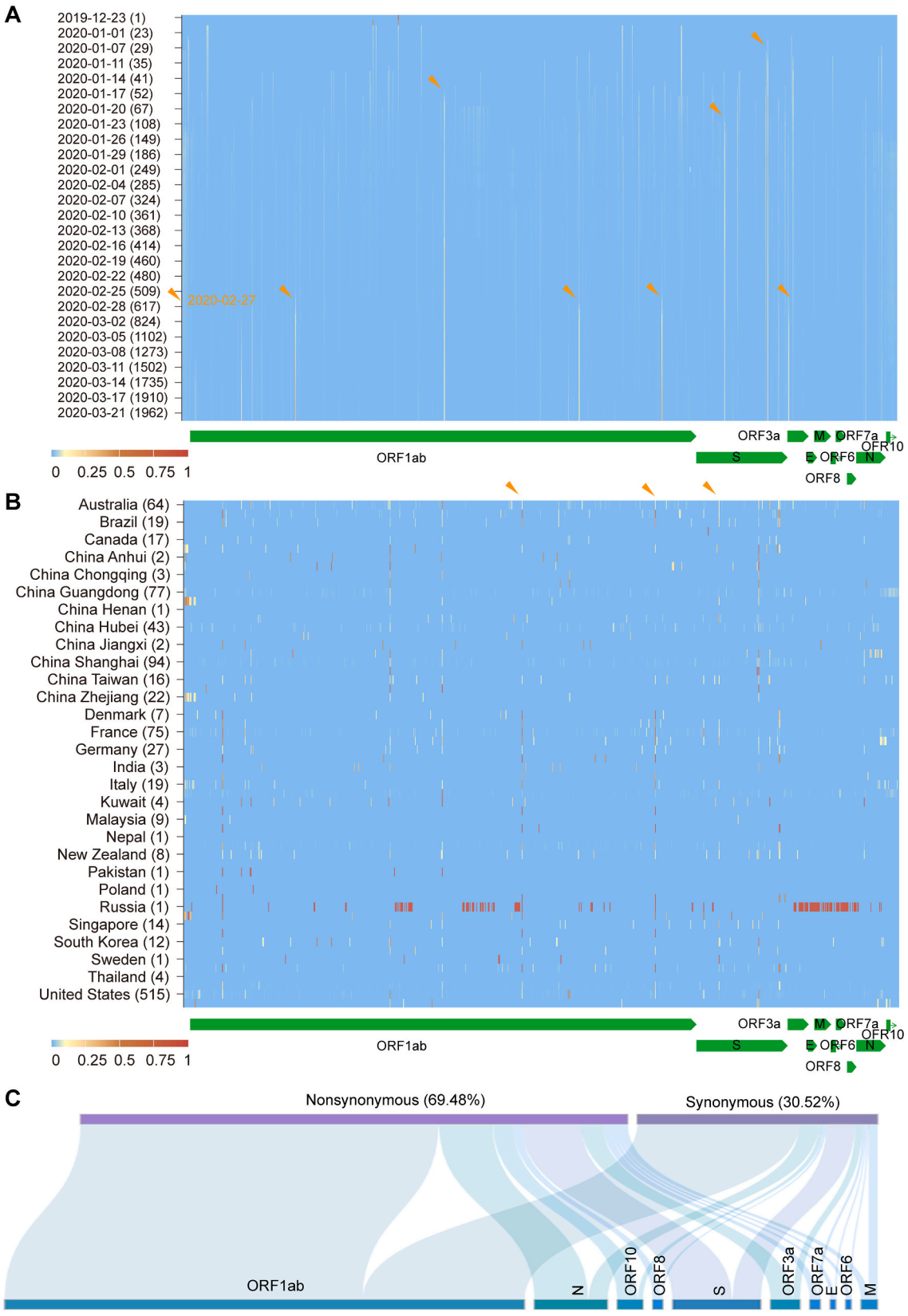

Fig. 1. Mutant landscape of genetic variations in SARS-CoV-2 strains. Genetic variation heatmap of SARS-CoV-2 strains over time (A) and country (B). Strains from different time points and different countries (partial) are indicated in the figure. Each vertical line shows mutation loci, yellow arrows indicate significant genetic variations. (C) Sankey diagram of coding regions in SARS-CoV-2 strains. Nonsynonymous and synonymous variations are shown on top, each gene region of SARS-CoV-2 is indicated on bottom.

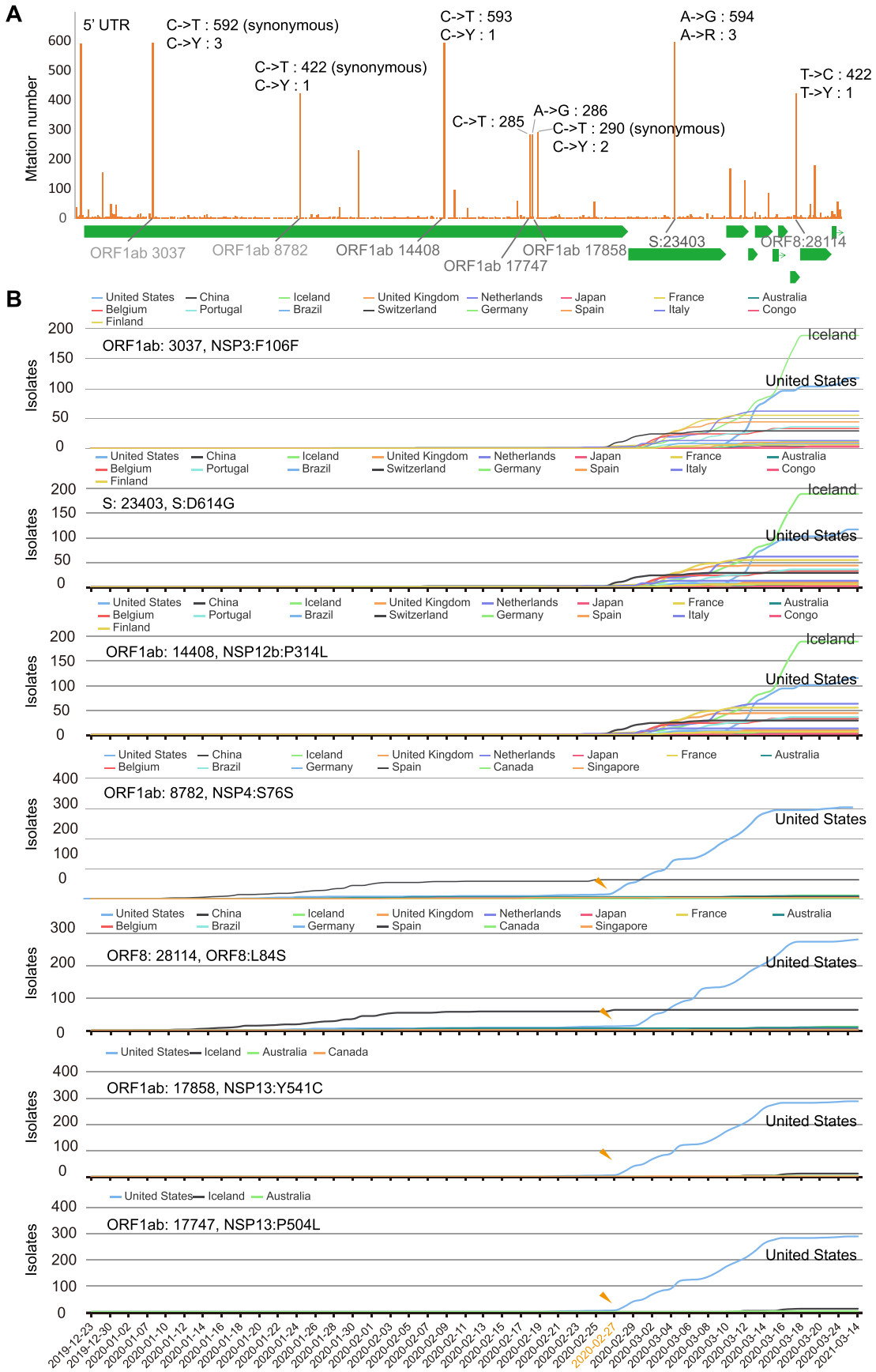

Fig. 2. Characteristics of genetic variations in SARS-CoV-2 strains. (A) Number of strains at each locus of the genetic variations. All gene regions of SARS-CoV-2 are complied in histogram, with top loci labeled. (B) Variation dynamic curve for occurrence of genetic variations in different countries. Significant loci, ORF1ab: 3037, S: 23403, ORF1ab: 14408, ORF1ab: 8782, ORF8: 28114, ORF1ab: 17858, ORF1ab: 17747, are displayed by detected time and country.

diverse intra-host populations (Domingo et al., 2012), NGS data was used to produce thousands to millions of reads from a mixed sample and to estimate viral quasispecies (Giallonardo et al., 2014). Here, the haplotype maps of SARS-CoV-2 strains across countries were based on genetic variations from high-quality genome sequencing data. From the 2250 viral strains, we identified 1210 haplotypes in total. Importantly, we observed that haplotype (H (140)) occurred in more strains than any other haplotype. Furthermore, 97.14% (136/140) of these strains were from the United States, with the remaining 2.85% (3/140) from Canada, thus showing significant differences between the two countries (P = 7.90E-14; Fig. 3A). These results indicate that this virus strain from the United States exhibited the aggregation in the population than any other strain.

As different viruses follow different patterns of variation, phylogenetic trees can be used to investigate their variation features (Holmes, 2008). Here, we constructed a phylogenetic tree of SARS-CoV-2 strains from the 2019nCoVR dataset using default settings (Zhao et al., Song et al., 2020). Results identified a clade enriched with strains from the United States (enriched clade, Fig. 3B). Furthermore, 236 strains fell within the enriched cluster, while 140 strains belong to Haplotype H (140) and the enriched cluster, the strains within haplotype H (140) coinciding with the strains in the enriched clade (Fig. 3C).

3.4. ORF1ab variations (NSP13: P504L and NSP13: Y541C) in SARS-CoV-2

We next analyzed the clade of United States strains in the phylogenetic tree. Results showed that amino acid variation in ORF1ab (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C) was an important characteristic of the clade (Fig. 4A). From previous analysis, genetic variations in ORF1ab: 17858 (A- > G, 286 strains) and ORF1ab: 17747 (C- > T, 285 strains) corresponded to amino acid variations in ORF1ab: p.5865Y > C (NSP13: Y541C) and ORF1ab: p.5828P > L (NSP13: P504L). Among the 286 strains, 285 strains contained amino acid variations in ORF1ab at the 5865 and 5828 sites, and one strain from Canada contained variation in ORF1ab at site 5865. In addition, 95% of variants were from the United States, 4% were from Australia, and 1% were from Iceland (Fig. 4B). When comparing the differences in ORF1ab variations at amino acid sites 5828 and 5865, we found significant differences among these three countries (P = 2.92E-66, Fig. 4C).

We also found that 271 variants (NSP13: P504L and NSP13: Y541C) from the United States, accounting for 51.13% of all strains from this country, contained the ORF1ab variants. In addition, most of them were from Washington State, or its counties and cities, with only three from Wisconsin, four from California, five from Utah, six from Minnesota, and 12 from unreported sources in the United States (Fig. 4D). Importantly, we found significant differences (P = 1.56E-23) among the states of Washington, Utah, Minnesota, Wisconsin, and California (Fig. 4E). Furthermore, 69.05% of strains from Washington State showed variation in ORF1ab (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C), with some regions showing a 100% variation rate, e.g., Umatilla County, Snohomish County, Clark County, and Tacoma (Fig. 4E).

3.5. COVID-19 outcomes in states with different proportions of strains containing variation in NSP13: P504L and NSP13: Y541C

To further analyze the variation in ORF1ab (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C) in SARS-CoV-2, we explored variation frequency over time. Results demonstrated that variation frequency increased on 27 February 2020 and peaked on 15–16 March 2020 (Fig. 5A). We then calculated the infection, fatality, and recovery rates of COVID-19 in the United States. On 15 and 16 March 2020, infections increased by more than 1000 cases, totaling 2234 and 3487 cases, respectively. In addition, the recovery rate dropped on 27 February 2020 and the fatality rate increased on 28 February 2020 (Fig. 5B). These date links between ORF1ab variation and COVID-19 condition were not found in Iceland or Australia (Fig. S4A, B). Furthermore, the number of infections in countries other than China exceeded the number of infections in China on 15 March 2020, with the fatality rate exceeded on 17 March 2020 (Fig. 5C).

By the detection time on 2 April 2020, Washington State had one of the highest fatality rates in the United States, despite not having a particularly high infection number (Fig. 5D). Furthermore, when comparing COVID-19 infections and deaths between Washington State and Minnesota, Utah, Wisconsin, and California, which contained a small percentage of strains with the ORF1ab variations (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C), significant differences were found (P = 3.64E-29), including in the death and recovery rates (P = 4.12E-21, Fig. 5E). In addition, King County, Snohomish County, Pierce County, Clark County, and Grant County, with more than half of strains containing ORF1ab variation (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C), showed high infection and fatality rates (Fig. 5F). Moreover, by the latest data on 18 May 2020, the fatality rate of Washington State was still higher than Minnesota, Utah, Wisconsin, and California, and showed significant differences among them (P = 4.4688E-53, Fig. S5A—C).

3.6. Effects of NSP13: P504L and NSP13: Y541C on NSP13 function

To better understand the consequences of amino acid variation in NSP13: P504L and NSP13: Y541C in NSP13 of SARS-CoV-2 (Jia et al., 2019), we performed sequence alignment of the variant against other coronaviruses. Of the 16 coronaviruses, the ORF1ab amino acids at sites 5828 and 5865 were conserved, with P (proline) at site 5828 and Y (tyrosine) at site 5865 (Fig. 6A). However, in the United States strains with ORF1ab variations, we found that the amino acids at sites 5828 and 5865 mutated into L (leucine) and C (cystine), respectively, which resulted in changes in the α-helix, β-sheet, and β-turn secondary structures of the protein near these sites (Fig. 6B). After modeling the protein structure, the variants showed different global quality scores in QMEAN, Cβ, all-atom, solvation, and torsion (Fig. S6).

Also, we used prediction tools based on computational methods to predict whether variation (P504L and Y541C) influenced the NSP13 function. The PolyPhen-2 tool showed high scores of the variants (NSP13: P504L and NSP13: Y541C) in HumDiv and HumVar, with the tendency that the closer the score is to 1.0, the greater effect on protein function (Fig. 6C). Alignment of 415 protein sequences with 30 clusters in PROVEAN v1.1 (Supplementary Tables S2–3) showed a deleterious effect on NSP13 (Fig. 6D). These results demonstrated that the variations (NSP13: P504L and NSP13: Y541C) tended to change the function of the NSP13.

4. Discussion

Coronaviruses, such as SARS-CoV, MERS-CoV, and now SARS-CoV-2, cause severe disease and pose a major threat to human health (Hilgenfeld and Peiris, 2013). The fatality rates of SARS and MERS are around 10% and 34%, respectively (Meo et al., 2020), whereas that of SARS-CoV-2 is about 2.22%. However, despite the relatively low fatality rate, SARS-CoV-2 has caused more deaths than SARS and MERS combined. It has been reported that SARS-CoV-2 accumulates an average of one or two mutations per month (Andersen et al., 2020) and an average of 7.23 mutations per sample (Mercatelli and Giorgi, 2020). Studies on mutation analysis of SARS-CoV-2 showed that NSP13: Y541C and NSP13: P504L were the top mutations in sequenced SARS-CoV-2 genomes (Mercatelli and Giorgi, 2020; Toyoshima et al., 2020), which were consistent with the NSP13 variants identified by our study.

As the branches of a viral phylogenetic tree can explain viral dynamics across the globe (Holmes, 2008), we found that strains containing p.5828P > L and p.5865Y > C (NSP13: P504L and NSP13:

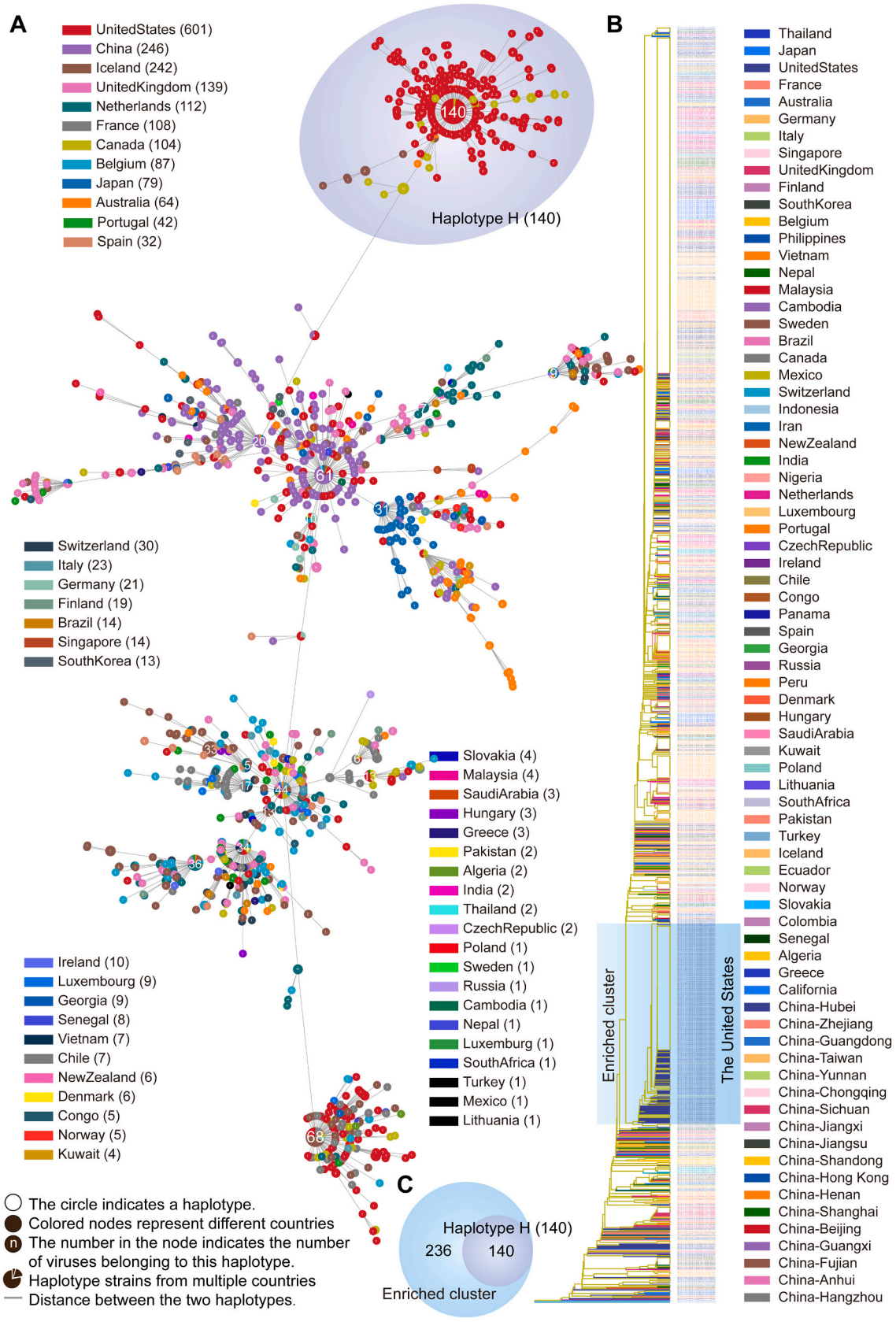

Fig. 3. Variation characteristics of SARS-CoV-2 strains. (A) Viral genotyping maps for haplotypes by country; gray circle indicates haplotype H (140). Strains from different countries are indicated in the figure, purple circle indicates haplotype type containing ORF1ab variation in 17,747 and 17,858 loci. (B) Phylogenetic tree across countries based on genome variation of SARS-CoV-2 strains downloaded from the 2019nCoVR dataset with default settings. Strains from different countries are indicated in the figure, blue square indicates enriched clade. (C) Venn diagram of strains with haplotype H (140) and enriched clade. Number of merged strains = 140.

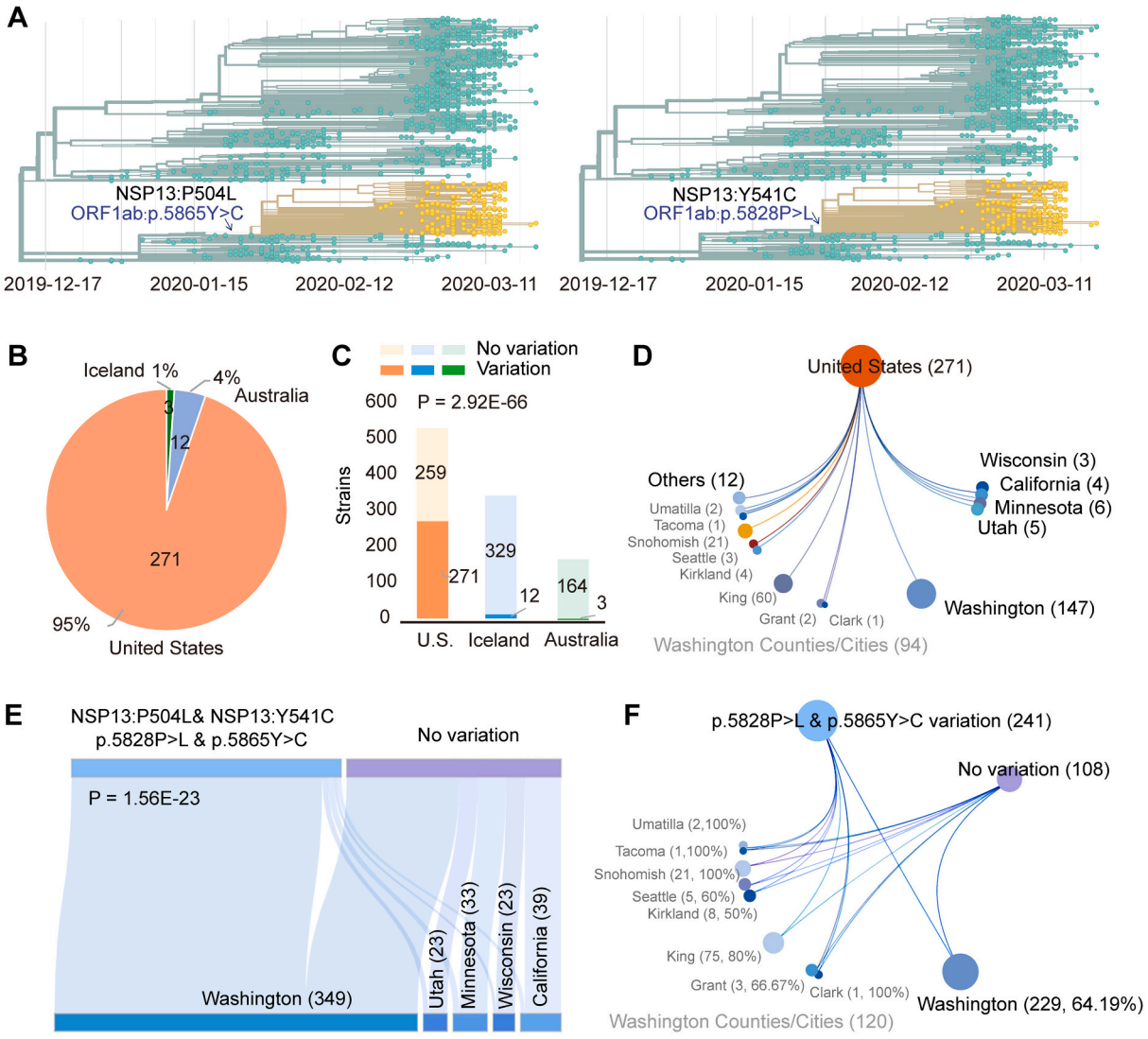

Fig. 4. ORF1ab variations (NSP13: P504L and NSP13: Y541C) in SARS-CoV-2. (A) Phylogenetic tree of strains downloaded from Nextstrain. Strains are colored by amino acid variations in ORF1ab (p.5828P > L or p.5865Y > C) (yellow) or not. (B) Pie chart of variants from Iceland, Australia, and the United States. (C) Column chart of variants from the United States, Iceland, and Australia, with number of strains labeled. (D) Distribution of 271 variants in ORF1ab (p.5828P > L and p.5865Y > C). (E) Sankey diagram of strains from Washington, Utah, Minnesota, Wisconsin, and California states, with strain number of each state labeled. (F) Distribution of variants in ORF1ab (p.5828P > L and p.5865Y > C) or not in Washington State.

Y541C) were responsible for the phylogenetic branches. Toyoshima et al. identified three clusters with a different mortality rate of COVID-19. In particular, the ORF1ab 4715 L and S protein 614G variants were correlated with a higher fatality rate. Although the NSP13: P505L and NSP13: Y541C variants were identified in Cluster 3, mainly in the United States, Australia, and Canada, the fatality rate of Cluster 3 showed no statistical significance compared with Cluster 1 and Cluster 2 (Toyoshima et al., 2020).

Using the genomic diversity of mutations in early SARS-CoV-2 strains, scientists have identified two distinct mutations, i.e., S or L type, with the S type considered more aggressive and faster spreading (Tang et al., 2020). From the analysis of early trends in the COVID-19 evolutionary patterns, Velazquez-Salinas et al. for the first time identified the effect of mutation ORF1ab:5865 as a factor of phylogenetic divergence of SARS-CoV-2 (Velazquez-Salinas et al., 2020). Our analysis of genetic variation in early SARS-CoV-2 strains found that most variations were in ORF1ab, and most variations in ORF1ab were non-synonymous. As nonsynonymous variations are usually under stronger negative selection than synonymous variations (Fusaro et al., 2011). It has been reported that functional constraints on viral genomes are weakened after the disruption of ORF1ab in SARS-CoV-2 (Tang et al., 2020). Moreover, in silico structure modeling of Nsp13 and Nsp14, potential dual-target inhibitors of SARS-CoV-2 with high binding affinity were identified (Gurung, 2020).

The data was identified and integrated from the 2019nCoVR database, with the non-redundant and duplicated biases (Zhao et al., 2020; Song et al., 2020). Factors associated with the fatality rate of COVID-19 fall into objective factors, such as the quality and capacity of the healthcare system, and subjective factors related to individual patients, such as age and history of chronic respiratory disease, hypertension, diabetes, and coronary heart disease (Zhou et al., 2020a). In Washington State, although the number of infections ranks tenth, the fatality rate ranks second (as of 2 April 2020), with a higher fatality rate than even New York. However, we found that for mortality rate in Washington State was high, with 69.05% of strains containing ORF1ab variants (NSP13: P504L and NSP13: Y541C), whereas other states (e.g., Minnesota, Utah, Wisconsin, and California) that only contained a small proportion of strains with the ORF1ab variants, showed better disease outcomes. However, outcomes in hosts infected with strains containing whether ORF1ab variations (NSP13: P504L and NSP13: Y541C) or not in

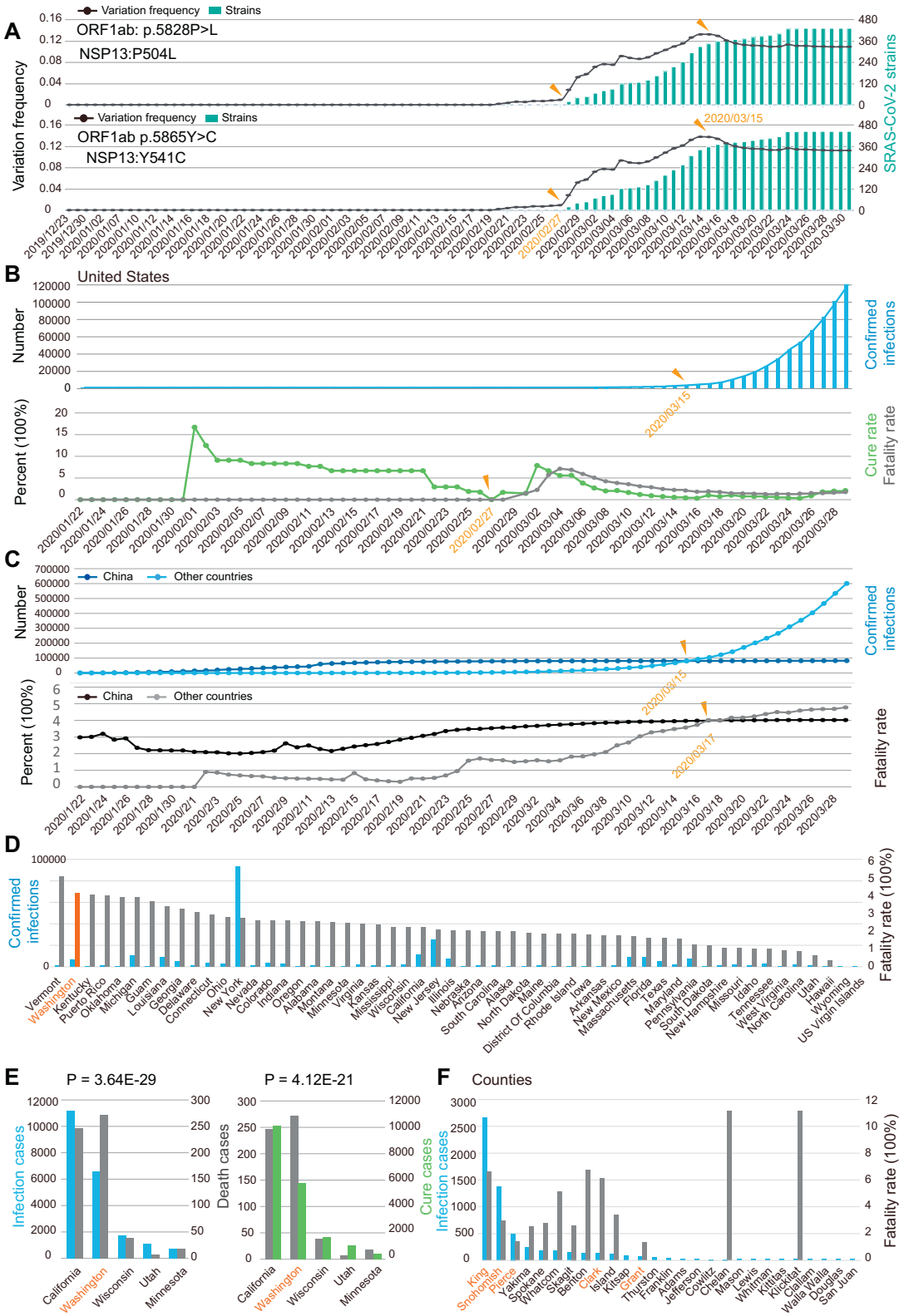

Fig. 5. COVID-19 outcomes in states with different proportions of strains containing variation in NSP13: P504L and NSP13: Y541C. (A) Variation frequency curve of ORF1ab p.5828P > L (NSP13: P504L) and ORF1ab p.5865Y > C (NSP13: Y541C), black line chart indicates variation frequency, green histogram indicates SARS-CoV-2 strains. (B) Infection, recovery, and fatality rates in the United States, blue histogram indicates infection cases, green line chart indicates recovery rate, and gray line chart indicates fatality rate. (C) Infection, recovery, and fatality rates in China and other countries. (D) Infection and fatality rates in different states in the United States, blue histogram indicates infection cases, gray histogram indicates fatality rate. (E) Infection, fatality, and recovery rates in different states, blue histogram indicates infection cases, gray histogram indicates fatality rate, and green histogram indicates recovery rate. (F) Infection and fatality rates in Washington State counties.

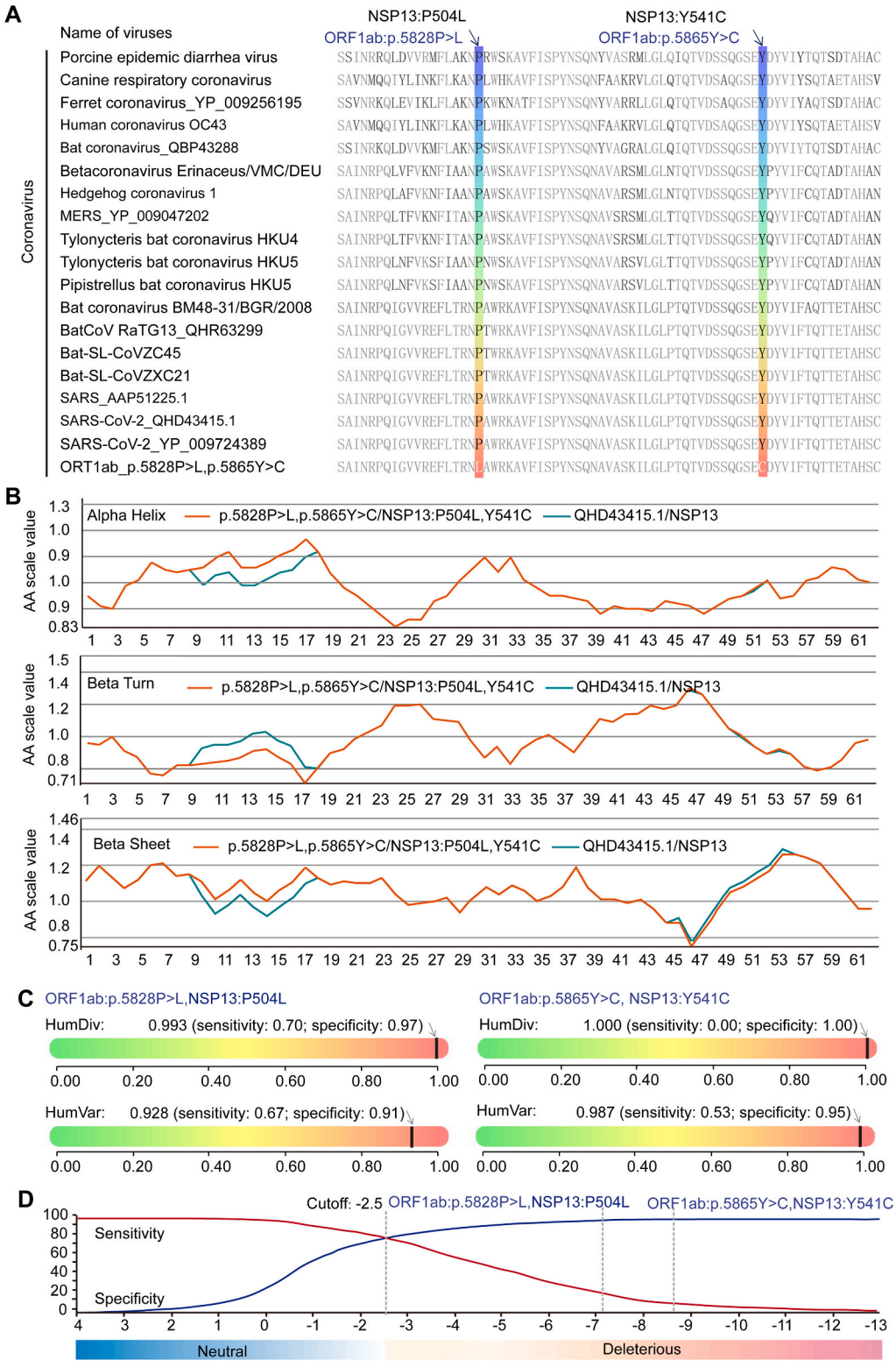

Fig. 6. Effects of NSP13: P504L and NSP13: Y541C on NSP13 function. (A) Sequence alignment of SARS-CoV-2 and other coronavirus, with amino acid sequences around ORF1ab variations aligned. (B) Results of protein secondary structure prediction in ORF1ab variations (QHD43415.1: p.5828P > L and p.5865Y > C) and ORF1ab (QHD43415.1), with α-helix, β-sheet, and β-turn structures of variant and QHD43415.1 displayed. Prediction results of ORF1ab variations (NSP13: P504L and NSP13: Y541C) based on PolyPhen-2 (C), and PROVEAN v1.1 (D).

the same States were missing, making it difficult to directly build the correlation between them.

The only way to know whether a nonsynonymous variation of a virus affects its function is to study it in cell assays or animal models to clarify entrance and transmission processes (Muth et al., 2018). As Velazquez-Salinas's study identified that residue 5865 (NSP13: Y541C) was under directional selection, which is also shown by the datamonkey evolutionary server, the variants (NSP13: P504L and NSP13: Y541C) may be under experimenting positive selection (Velazquez-Salinas et al., 2020). Besides, we only predicted secondary structures of proteins (α-helix, β-sheet, and β-turn) and functional effects based on computational methods, without the application of cell cultures or animal models to demonstrate the consequences of the variations, which need further study. Furthermore, virus strains carrying the two key variations were

also found in Iceland and Australia but were not dominant in these two countries. The mutation frequencies of these two loci were high in Mexico, Canada, and the United States (Fig. S7), which are geographically close to each other. And genetic mutation heatmaps of regions showed that the mutation type of Iceland, Australia, and the United States were different (Fig. S8).

In this research, we explored 1962 high-quality genomes of early SARS-CoV-2 to identify the changing trends in genetic variations over time and by country. Haplotype mapping and phylogenetic tree analysis showed that amino acid variations in ORF1ab (p.5828P > L and p.5865Y > C, NSP13: P504L and NSP13: Y541C) were important characteristics of this clade. Moreover, different disease outcomes were found among states containing different proportions of the variants with NSP13 variations (at 501 and 541 loci) in the United States, especially in Washington State. Thus, by analyzing genetic variations in SARS-CoV-2, we identified a correlation between amino acid variation in ORF1ab and COVID-19 outcomes.

Data availability statement

The sequences of the SARS-CoV-2 strains used in the analysis are available, upon free registration, from the GISAID database (https://www.gisaid.org/). The 2019 Novel Coronavirus Resource (2019nCoVR) database (v2.1) are available (https://bigd.big.ac.cn/ncov).

Author contribution

C.C. performed the bioinformatics analyses and wrote the manuscript. P.W. and G.L. designed the study. L.H., Y.T., Y.Q., H.S., and W.D. provided critical feedback on the experiments as well as discussions on and coordination of the project.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the Research-Oriented Clinician Funding Program of Tongji Medical College, Huazhong University of Science and Technology. We are grateful for the 2019nCoVR (v2.1) database and the computational analysis tools of Nextstrain, PROVEAN (v1.1), and PolyPhen-2. We sincerely thank the work done by all anti-epidemic personnel.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.104831.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., et al., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7 (4), 248–249. https://doi.org/10.1038/nmeth0410-248.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26 (4), 450–452. https://doi.org/10.1038/s41591-020-0820-9.

Becerra-Flores, M., Cardozo, T., 2020. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int. J. Clin. Pract. 74 (8), e13525 https://doi.org/10.1111/ijcp.13525.

Bienert, S., Waterhouse, A., de Beer, T.A., Tauriello, G., Studer, G., Bordoli, L., et al., 2017. The SWISS-MODEL repository-new features and functionality. Nucleic Acids Res. 45 (D1), D313–D319. https://doi.org/10.1093/nar/gkw1132.

Canhui, C., Huang, L., Liu, K., Ma, K., Tian, Y., Qin, Y., et al., 2020. Amino acid variation analysis of surface spike glycoprotein at 614 in SARS-CoV-2 strains. Gen. Dis. 4 (4), 567–577. https://doi.org/10.1016/j.gendis.2020.05.006.

Choi, Y., Chan, A.P., 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31 (16), 2745–2747. https://doi.org/10.1093/bioinformatics/btv195.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7 (10), e46688. https://doi.org/10.1371/journal.pone.0046688.

Chou, P.Y., Fasman, G.D., 1974. Prediction of protein conformation. Biochemistry 13 (2), 222–245. https://doi.org/10.1021/bi00699a002.

Domingo, E., Sheldon, J., Perales, C., 2012. Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. 76 (2), 159–216. https://doi.org/10.1128/MMBR.05023-11.

du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., et al., 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 371 (6530), 708–712. https://doi.org/10.1126/science.abf2946.

Fusaro, A., Monne, I., Salviato, A., Valastro, V., Schivo, A., Amarin, N.M., et al., 2011. Phylogeography and evolutionary history of reassortant H9N2 viruses with potential human health implications. J. Virol. 85 (16), 8413–8421. https://doi.org/10.1128/JVI.00219-11.

Giallonardo, F.D., Topfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., et al., 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. 42 (14), e115 https://doi.org/10.1093/nar/gku537.

Gurung, A.B., 2020. In silico structure modelling of SARS-CoV-2 Nsp13 helicase and Nsp14 and repurposing of FDA approved antiviral drugs as dual inhibitors. Gene Rep. 21, 100860. https://doi.org/10.1016/j.genrep.2020.100860.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., et al., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34 (23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.

Hilgenfeld, R., Peiris, M., 2013. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. Antivir. Res. 100 (1), 286–295. https://doi.org/10.1016/j.antiviral.2013.08.015.

Holmes, E.C., 2008. Evolutionary history and phylogeography of human viruses. Annu. Rev. Microbiol. 62, 307–328. https://doi.org/10.1146/annurev.micro.62.081307.162912.

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., et al., 2018. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerg Microbes Infect 7 (1), 154. https://doi.org/10.1038/s41426-018-0155-5.

Jia, Z., Yan, L., Ren, Z., Wu, L., Wang, J., Guo, J., et al., 2019. Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. Nucleic Acids Res. 47 (12), 6538–6550. https://doi.org/10.1093/nar/gkz409.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al., 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182 (4), 812–827 e819. https://doi.org/10.1016/j.cell.2020.06.043.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395 (10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.

Meo, S.A., Alhowikan, A.M., Al-Khlaiwi, T., Meo, I.M., Halepoto, D.M., Iqbal, M., et al., 2020. Novel coronavirus 2019-nCoV: prevalence, biological and clinical characteristics comparison with SARS-CoV and MERS-CoV. Eur. Rev. Med. Pharmacol. Sci. 24 (4), 2012–2019. https://doi.org/10.26355/eurrev_202002_20379.

Mercatelli, D., Giorgi, F.M., 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. 11, 1800. https://doi.org/10.3389/fmicb.2020.01800.

Muth, D., Corman, V.M., Roth, H., Binger, T., Dijkman, R., Gottula, L.T., et al., 2018. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci. Rep. 8 (1), 15177. https://doi.org/10.1038/s41598-018-33487-8.

Ogando, N.S., Zevenhoven-Dobbe, J.C., van der Meer, Y., Bredenbeek, P.J., Posthuma, C. C., Snijder, E.J., 2020. The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2. J. Virol. 94 (23) https://doi.org/10.1128/JVI.01246-20.

Sawicki, S.G., Sawicki, D.L., Siddell, S.G., 2007. A contemporary view of coronavirus transcription. J. Virol. 81 (1), 20–29. https://doi.org/10.1128/JVI.01358-06.

Song, S., Ma, L., Zou, D., Tian, D., Li, C., Zhu, J., et al., 2020. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics, Proteomics & Bioinformatics.

Studer, G., Tauriello, G., Bienert, S., Biasini, M., Johner, N., Schwede, T., 2021. ProMod3-A versatile homology modelling toolbox. PLoS Comput. Biol. 17 (1), e1008667 https://doi.org/10.1371/journal.pcbi.1008667.

Tang, X.L., Wu, C., Li, X., Song, Y.H., Yao, X.M., Wu, X.K., et al., 2020. On the origin and continuing evolution of SARS-CoV-National Science Review 2020. https://doi.org/10.1093/nsr/nwaa036 nwaa036.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. 65 (12), 1075–1082. https://doi.org/10.1038/s10038-020-0808-9.

Velazquez-Salinas, L., Zarate, S., Eberl, S., Gladue, D.P., Novella, I., Borca, M.V., 2020. Positive selection of ORF1ab, ORF3a, and ORF8 genes drives the early evolutionary trends of SARS-CoV-2 during the 2020 COVID-19 pandemic. Front. Microbiol. 11, 550674. https://doi.org/10.3389/fmicb.2020.550674.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46 (W1), W296–W303. https://doi.org/10.1093/nar/gky427.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3.

Zhao, W.M., Song, S.H., Chen, M.L., Zou, D., Ma, L.N., Ma, Y.K., et al., 2020. The 2019 novel coronavirus resource. Yi Chuan 42 (2), 212–221. https://doi.org/10.16288/j.yczz.20-030.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al., 2020a. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 395 (10229), 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., et al., 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al., 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382 (8), 727–733. https://doi.org/10.1056/NEJMoa2001017.